Special Report
SR-00-39

September, 2000

Third-Party Checking of 2000 Scaling and Linking for the Kentucky Core Content Test

R. Gene Hoffman
Arthur A. Thacker

Prepared for:

The Kentucky Department of Education
500 Mero Street
Frankfort, KY 40601

Contract Number BP 010492

Human Resources Research Organization
66 Canal Center Plaza, Suite 400 • Alexandria, VA 22314-1591

Third-Party Checking of 2000 Scaling and Linking for the Kentucky Core Content Test

Table of Contents

Summary

CTB and HumRRO independently calculated the scaled/linked raw-score-to-scale-score tables for the 2000 Kentucky Core Content Test. From those tables, both identified cutpoints that could be used for assigning student performance classifications and later converted to school accountability indexes. Differences between the calculations of CTB and those of HumRRO were small and did not affect student classification decisions in any instance. Very slight differences in item parameters, likely due to rounding procedures, did not yield any differences in raw-score-to-scale-score tables or cutpoints. Given that our scaling and linking results are identical with those of CTB, we can be assured that CTB did not commit processing errors.

Third-Party Checking of 2000 Scaling and Linking for the Kentucky Core Content Test

Introduction

In order to make the transition from the KIRIS test to the Kentucky Core Content Test with the minimum amount of disruption, a system of linking the old test with the new was necessarily devised. This link allowed Kentucky to maintain consistency in its student performance levels and to apply the student Kentucky Core Content Test scores to a newly revised accountability calculation. The main difficulty in linking the two tests was that KIRIS only applied student scores on the open-response section of the test toward a school's accountability index and toward individual student performance levels. The Kentucky Core Content Test uses both open-response and multiple-choice format questions to make those determinations. Students still receive ratings in terms of the Novice, Apprentice, Proficient, and Distinguished levels of performance, but multiple-choice questions are now included in those determinations. A two-step process was used to make the link from the Kentucky Core Content Test back to the KIRIS scale on which student performance standards had been set in 1993 (Kentucky Department of Education, 1997). The first step involved analysis of 1998 data in which multiple-choice and open-response items were combined on a single scale and that combined scale equated to the open-response-only scale. HumRRO, in an earlier report (Hoffman, Thacker, & McBride, 1999), performed a third-party evaluation of those procedures. The second step, for linking the Kentucky Core Content Test back to the KIRIS scale, was to link the 1999 test data to the newly created combined scale. HumRRO also performed a third-party evaluation of those procedures (Hoffman & Thacker, 1999).

The 2000 administration of the Kentucky Core Content Test was also linked back to the combined scale. This was accomplished by linking the 2000 test back to the 1999 test. The procedures for doing so mimic the procedures used in 1999. This report represents HumRRO's third-party check of the scaling and linking of the 2000 Kentucky Core Content Test.

Scaling and Linking Procedures

Item data from all forms were scaled using CTB's PARDUX program. Item parameters were then divided by form and entered into CTB's FLUX program to create raw-score-to-scale-score conversion tables. The scaling process included adjusting item parameters by PARDUX application of the Stocking-Lord procedure to items linking the 2000 Kentucky Core Content Test to the 1999 administration of the test. One form from each grade/subject was identified from the 1999 Kentucky Core Content Test to serve as an anchor form. Each anchor form was readministered in 2000 with all items intact and occurring in the same sequence as in 1999. All anchor item parameters come from the multiple-choice items included on the anchor form. Open-response items were repeated on the anchor forms for form construction consistency and to ensure that contextual clues that may have been present in 1999 were repeated for 2000. Cutpoints established in 1993 could then be applied to the 2000 scaling results. For Arts & Humanities, Practical Living/Vocational Studies, and Grade 10 Reading, cutpoints were established by equipercentile equating to 1998 performance level distributions in 1999. 2000 results were linked to the scale established in 1999 for those subjects.

Scope of Third-Party Checking

HumRRO conducted parallel analyses to accomplish scaling and linking for the 2000 data. Because of the severe time limits, HumRRO's analyses were constrained in two ways. First, CTB selected the calibration sample based on criteria set by KDE. HumRRO did not independently select a sample, but rather used the CTB selection. Second, CTB conducted item-total raw score correlations on multiple-choice data and identified items to be excluded from scaling because of sufficiently poor item biserial correlations. HumRRO did not independently recalculate these correlations. Any such aberation, however, would be detected in Pardux scaling runs.

Processing Steps

HumRRO took the following steps for each grade/subject tested:

1. Create anchor files (PARDUX *.anc) of multiple-choice test items that appear on the anchor form. These anchor items are used to link the 2000 test to the 1999 scale which was previously adjusted to the 1993 scale. A special SAS program was written to create this file.

2. Prepare control files (PARDUX *.ctl) which contain the constraints used for item parameter estimation, student proficiency estimation, maximum number of items, etc. The SAS program used to create anchor files included a routine to print out a control file.

3. Create working files (PARDUX *.RWO) from the 2000 Kentucky Core Content Test calibration sample. These files include both open-response and multiple-choice data. Two different SAS programs were used to create *.RWO files. Because item placement in the *.RWO file is tedious, one program was used to generate lines of code that moved input data into the correct output location. The second program applied those lines of code to the input files.

4. Estimate parameters for Kentucky Core Content Test items using PARDUX.

5. Perform Stocking-Lord transformation using PARDUX. The results of this transformation include a slope and intercept constant for linking the 2000 Kentucky Core Content Test back to 1999.

6. Confirm that the equating constants from Step 5 match those derived by CTB.

7. Create parameter files (FLUX *.par) for each test form for use in preparation of raw-score-to-scale-score tables. This was a "cut and paste" word processing task using PARDUX output of item parameters from step 4.

8. Create files (FLUX *.hlk) containing the scale limits (325 and 800) and constants from the Stocking-Lord transformation. This was a simple word processing task.

9. Create raw-score-to-scale-score transformation tables for each form using FLUX.

10. Confirm that the raw-score-to-scale-score transformation tables from Step 9 match those derived by CTB.

11. Confirm that the cutpoints set by CTB were consistent with established cutpoints from the KDE (1997) Cycle 2 Technical Manual and Wise (1998) Grade Shift Report.

Results

After performing periodic checks with CTB as individual tests were scaled and equated, HumRRO and CTB reached appropriate levels of agreement on the equating constants for all grade/subjects. The agreement level was typically very close, as indicated by Table 1. Very small differences in M1 and M2 (slope and intercept constants) results occurred, but did not affect the raw-score-to-scale-score tables and hence did not affect any student's performance classification.

Table 1 summarizes the results of this study. It identifies the grade and subject for each test in the first two columns. The third column identifies problem items and references the solutions that were reached by CTB and verified by HumRRO. The next four columns contain the M1 and M2 (slope and intercept) constants obtained from the Stocking-Lord transformation. HumRRO computed the first set of constants, CTB the second. The seventh and eighth columns contain the difference between the M1 and M2 constants computed by HumRRO and those computed by CTB. As can be seen from these columns, HumRRO and CTB are in very close agreement for all grade/subjects.

The last column in Table 1 is a verification of the exact agreement between CTB and HumRRO for the raw-score-to-scale-score tables. Cutpoints from to those tables are used to assign students to performance categories (Novice Apprentice, Proficient, or Distinguished), that are in turn used in the computation of each school's accountability index. CTB and HumRRO were in exact agreement for all raw-score-to-scale-score tables for every grade/subject.

The asterisks from the third column of Table 1 represent problem items. Each asterisk is referenced with the specific problem that occurred and the solution. All problem items were dealt with during the parameter estimation phase of the scaling and equating process. No item for which parameters were estimated was eliminated from the Stocking-Lord procedure. The same column indicates whether or not convergence was reached during parameter estimation. If convergence was not reached after 50 iterations by the Pardux program, the solution at stage 50 was accepted by mutual agreement.

HumRRO also verified the cutpoints on the raw-score-to-scale-score tables. Cutpoints were assigned using a previously agreed upon rounding rule. Each cutpoint was rounded normally and the result represented the first scale score for the next higher category. Hence, if the cutpoint between Novice and Apprentice was 500.4, a score of 500 would be designated Apprentice. If the cutpoint was 500.6, the same score would be designated Novice. Using this rule, all cutpoints for all forms of each subject test were identical for HumRRO and CTB. HumRRO verified cutpoints between Novice and Apprentice, between Apprentice and Proficient, and between Proficient and Distinguished performance categories.

Table 1
Comparison of HumRRO and CTB Scaling and Linking Results.

Grade	Subject	Problems	HumRRO M1	HumRRO M2	CTB M1	CTB M2	Difference M1	Difference M2	RS-SS Exact Agreement
04	RD	135[1]	31.11432	547.14490	31.11436	547.14496	0.00004	0.00006	Yes
04	SC	None	25.90414	543.42328	25.90418	543.42334	0.00004	0.00006	Yes
05	A&H	None	49.45951	506.50064	49.45951	506.50064	0.00000	0.00000	Yes
05	MA	None	34.94686	556.45612	34.94686	556.45605	0.00000	0.00007	Yes
05	PL	None (No Convergence)	47.12158	500.60931	47.12158	500.60931	0.00000	0.00000	Yes
05	SS	None	31.88767	537.79883	31.88767	537.79877	0.00000	0.00006	Yes
07	RD	61[2]	30.42912	510.97256	30.42929	510.97272	0.00017	0.00016	Yes
07	SC	None	25.55330	500.75342	25.55331	500.75345	0.00001	0.00003	Yes
08	A&H	None	47.87190	510.52655	47.87190	510.52655	0.00000	0.00000	Yes
08	MA	168[3] (No Convergence)	33.53253	530.76813	33.53255	530.76813	0.00002	0.00000	Yes
08	PL	None	43.54315	501.96674	43.54315	501.96674	0.00000	0.00000	Yes
08	SS	None	38.96319	510.24908	38.96318	510.24905	0.00001	0.00003	Yes
10	PL	110[4]	45.07659	503.44559	45.07659	503.44559	0.00000	0.00000	Yes
10	RD	None	50.03294	506.43399	50.03294	506.43399	0.00000	0.00000	Yes
11	A&H	5, 100, 110[5] (No Convergence)	47.41780	508.29224	47.41756	508.29236	0.00024	0.00012	Yes
11	MA	35, 130[6] (No Convergence)	40.46878	530.79681	40.46875	530.79663	0.00003	0.00018	Yes
11	SC	None	31.81342	541.73700	31.81334	541.73694	0.00008	0.00006	Yes
11	SS	None	46.59724	544.61591	46.59730	544.61580	0.00006	0.00011	Yes

[1] Grade 4 Reading: Item 135 did not estimate. Ran M-step and convergence was obtained.

[2] Grade 7 Reading: Item 61 was a non-discriminating item and was deleted. The item occurred on the anchor form and was also dropped last year.

[3] Grade 8 Mathematics: Item 168 negatively discriminated throughout most of the scale score range and was deleted.

[4] Grade 10 PL/VS: Item 110 did not estimate. Ran M-step and convergence was obtained.

[5] Grade 11 A&H: Item 100 was not scored. Item 110 did not estimate. Ran M-step and convergence was obtained. Item 5 was very flat throughout most of the scale score range and had an estimated b parameter of > 1000. Item 5 was deleted.

[6] Grade 11 Mathematics: Items 35 and 130 did not estimate. Ran M-step and convergence was obtained for item 130. Item 35 required M-step and hand fitting before convergence was reached.

Additional Anomalies

In addition to noting item problems, there are two form level anomalies that deserve mentioning. These anomalies in no way indicate a divergence between the results obtained by CTB and those obtained by HumRRO, but should be documented. The first regards form construction for the eighth grade social studies test. Each grade/subject test is composed of 12 forms. Each is numbered 1-6 and is designated either A or B. The Arts & Humanities and Practical Living/Vocational Studies tests, because of their limited number of items, are different for both scored items and pre-test items for each of the 12 forms. A and B forms of the tests in mathematics, reading, science, and social studies typically are identical except for the pretest items. The one exception to that rule is eighth grade social studies Form 5, for which A and B forms contain slightly different scored item sets. For that reason, the documentation contains separate Form 5A and 5B information. Separate raw-score-to-scale-score tables were created for these forms, and cutpoints were assigned and verified for each.

The other form level anomaly occurred in eleventh grade Arts & Humanities. One of the open-response questions (item 100 from Table 1) from Form 5B was not scored and was therefore not included in any analyses. Arts & Humanities forms only contain 10 scored items (2 open-response, 8 multiple-choice) totaling 24 possible raw score points (4 points for each open-response item, multiplied by 2 due to weighting, and 1 point each for multiple-choice). The elimination of one open-response item eliminated 8 points from the raw score scale. Several solutions were considered, including doubling the weight of the existing open-response question, assuming full credit on the missing item, etc. The final decision was to reduce the raw score scale and assign cutpoints normally. The raw-score-to-scale-score table for this form looks like the others, except that there are only 16 possible raw scores where the other forms have 24. This decision maintains the integrity of the scale and results in a distribution of students in each of the performance categories that is similar to other eleventh grade Arts & Humanities forms. The solution does mean, however, that for this form multiple-choice and open-response are weighted equally.

Documentation

To document the steps involved in scaling and linking the 1999 Kentucky Core Content Test we saved all electronic files used in data preparation, including SAS programs, SAS logs, and SAS output lists and all files produced during PARDUX scaling and FLUX transformations. These files have been submitted to KDE. Appendices from the 1999 report (Hoffman & Thacker, 1999) contain printed examples of important files that were submitted.

All electronic files submitted to KDE were zipped (Winzip) and are named according to the following code (where S = subject, G = grade level).

A. PARDUX Control File (SSGG00.CTL). This file contains the number of items, the maximum number of stages for PARDUX, the convergence criterion, parameter estimation limits, maximum and minimum values for proficiency estimates (theta), and other information. This file also contains information allowing the program to distinguish between open-response and multiple-choice items, the items to be calibrated, and the number of score levels for open-response data.

B. PARDUX Data File (SSGG00.RWO). This file contains the student score data. It is coded such that a 1 indicates a correct answer for a multiple-choice question and actual score levels (0-4) are recorded for student responses to open-response questions. To facilitate communication, HumRRO adhered to CTB's item order in constructing these data files.

C. PARDUX Anchor File (SSGG00.ANC). This file contains 1999 common-scaling item parameters for the 2000 Kentucky Core Content Test. These items were unchanged from 1999 to 2000. Only multiple-choice items are used in *.ANC files.

D. SAS Programs for Creating Anchor Files, PARDUX Control Files and *.RWO (Working Data) Files. The run logs for two different programs are included. The first program (see SSGGRWCD.log) assigns *.RWO locations for each item, creates anchor files from 1999 parameter files, creates control files, and writes lines of code that were then inserted into the second program. The second program (see SSGGrwo.log) merges multiple-choice raw data and open-response raw data to create an *.RWO file with items aligned according to CTB specifications.

E. PARDUX Parameter Estimation Summary (SSGG00SUM.TXT). This file provides a summary of the parameter estimation procedure run in PARDUX. It includes the limit data from the control file and also contains the number of stages PARDUX runs in order to reach convergence. It also contains the item numbers of items that could not be estimated and documents any items whose estimation reaches the maximum alpha parameter. This file identifies any problem items that might require additional manipulation before continuing the process.

F. PARDUX Parameter Estimation Details (SSGG00DET.TXT). This file is a thorough iteration of the item data during the final stage of parameter estimation.

G. PARDUX Parameter File (SSGG00.PAR). This file contains parameter estimates for all items designated by the *.CTL file. It is used for later data manipulation.

H. PARDUX TST File (SSGG00.tst). This file can be used to calculate form reliabilities. It was created and saved for each grade and subject tested. No form reliabilities were calculated for this project.

I. PARDUX VEC File (SSGG00.vec). This file contains all student data and includes an estimation of proficiency for each student's score data.

J. PARDUX Item Summaries Files, Status (SSGG00STAT.TXT). This file lists all items for a given test and their status after parameter estimation. Items are coded as either estimate OK, OK—default C, not estimated, or other codes. This file provides a different type of record for the parameter estimation.

K. PARDUX Item Summaries Files, Distribution (SSGG00DIST.TXT). This file contains the distribution of students who scored at each level on the open-response items. This file is useful for examining the way that scoring rubrics for these items operate and for ensuring that all open-response items have the correct number of functioning score levels.

L. PARDUX Item Summaries Files, Parameters (SSGG00PAR.TXT). This file contains the item parameters in more readily edited format than the *.PAR file. This file can easily be read into word processors and spreadsheet programs.

M. PARDUX Item Summaries Files, Standard Errors (SSGG00SE.TXT). This file contains the standard errors for each item including the errors for the various score levels on the open-response items.

N. PARDUX Item Summaries Files, FitQ1 (SSGG00Q1.TXT). This contains fit statistics for all items.

O. PARDUX Log File (SSGG00LOG.TXT). As each manipulation of data is completed, PARDUX maintains a log of the procedures and filenames. This log is saved in text format.

P. Stocking-Lord Plots (SSGG00_SL_PLOTS.doc). The Stocking-Lord transformation of the data, which provides the M1 and M2 values (slope and intercept) that allow for the later creation of scoring tables outputs three graphs (one each for the a, b, and c parameters and a fourth graph of P values) for each transformation. In this file the four graphs that result from the transformation using all anchor items are included. No anchor items were excluded during 2000 at this stage. This file also contains the Stocking-Lord log, a record of the Stocking-Lord procedure.

Q. FLUX control file (SSGG00.HLK). This file specifies the range of the scale scores as well as the M1 and M2 transformation constants to be used from the Stocking-Lord transformation.

R. FLUX Parameter Files by Form (SSGG00FORMX.PAR, etc., one for each Form). Each of the parameter files computed using PARDUX was divided to represent items from each test form. Typically, 30 items were scored from each form. Arts & Humanities and Practical Living/Vocational Studies forms contained 10 items to be scored.

S. Raw-score-to-scale-score Tables (SSGG00_RS_to_SS_Tables. A raw-score-to-scale-score table was produced for each form. These tables were saved as a single large Microsoft Word document for each tested grade/subject.

Conclusion

CTB and HumRRO independently calculated the scaled/linked raw-score-to-scale-score tables for the 2000 Kentucky Core Content Test. From those tables, both identified cutpoints that could be used for assigning student performance classifications and later converted to school accountability indexes. Differences between the calculations of CTB and those of HumRRO were small and did not affect student classification decisions in any instance. Given that our scaling and linking results are nearly identical with CTB, we can be assured that CTB did not commit processing errors.

References

Hoffman, R. G. & Thacker, A. A. (1999). *Third-party checking of 1999 scaling and linking for the Kentucky Core Content Test*. (HumRRO Report SP-WATSD-99-44). Alexandria, VA: Human Resources Research Organization.

Hoffman, R. G., Thacker, A. A., & McBride, J. R. (1999). *Documentation of third-party checking of 1998 pre-equating for Kentucky Core Content Test: IRT scaling of multiple choice and open response test items.* (HumRRO Report SP-WATSD-99-39). Alexandria, VA: Human Resources Research Organization.

Kentucky Department of Education (1997, May). KIRIS accountability cycle 2 technical manual: based on analysis of data from the 1992-93 through 1995-96 school years. Frankfort, KY: Author.